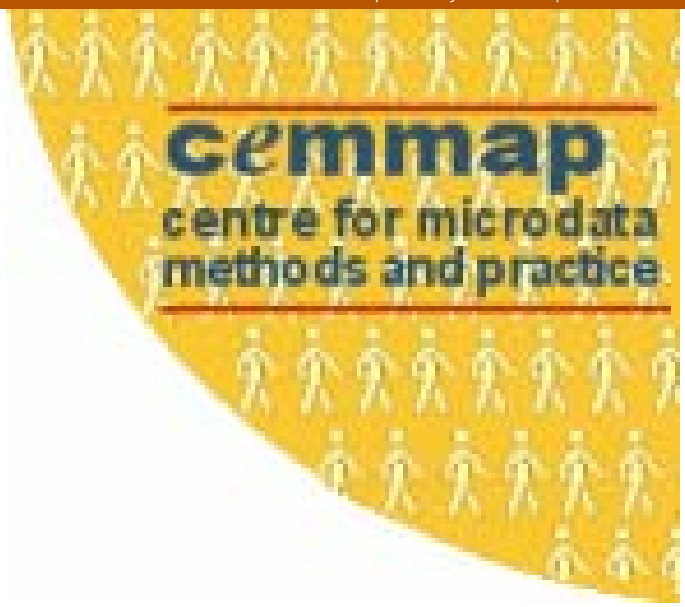

ESTIMATING A CLASS OF TRIANGULAR SIMULTANEOUS EQUATIONS MODELS WITHOUT EXCLUSION RESTRICTIONS

Roger Klein
Francis Vella

THE INSTITUTE FOR FISCAL STUDIES
DEPARTMENT OF ECONOMICS, UCL
cemmap working paper CWP08/05

Estimating a Class of Triangular Simultaneous Equations Models Without Exclusion Restrictions*

Roger Klein
Rutgers University

Francis Vella
European University Institute

July, 2005

Abstract

This paper provides a control function estimator to adjust for endogeneity in the triangular simultaneous equations model where there are no available exclusion restrictions to generate suitable instruments. Our approach is to exploit the dependence of the errors on exogenous variables (e.g. heteroscedasticity) to adjust the conventional control function estimator. The form of the error dependence on the exogenous variables is subject to restrictions, but is not parametrically specified. In addition to providing the estimator and deriving its large-sample properties, we present simulation evidence which indicates the estimator works well.

1 Introduction

Instrumental variables (IV) is a method commonly employed in empirical applications for estimating models with endogenous regressors. However,

*We are grateful to participants at numerous seminars over the past four years for various comments which have resulted in improvements to the paper. We would also like to thank Ethel Fonseca for helpful comments. We are particularly grateful to Whitney Newey for detailed comments which led to the current formulation of the problem. Any remaining errors are the sole responsibility of the authors.

while there is general agreement that IV is appropriate for a large class of models with endogeneity, there is frequently little agreement about the exclusion restrictions that this method typically requires in specific empirical applications. In fact, the difficulty in obtaining instruments has generated a rapidly growing and important literature related to inference in the presence of weak instruments (see, for example, Staiger and Stock 1999).

When the primary equation of interest contains an endogenous regressor, it is well known that IV is equivalent to an OLS regression that includes an additional regressor to control for endogeneity. Commonly, this additional variable or control is the reduced form residual for the endogenous regressor. In the linear case, as the control is a linear combination of the endogenous regressor and exogenous variables, the model is only identified in the presence of at least one exclusion restriction.[1]

In the above case the impact of the control is a constant that is estimated along with the parameters of interest. As a result, without further information, identification requires an exclusion restriction. However, when the error distribution depends on the exogenous variables, we show that it is possible and in some sense natural to develop a control whose impact is not constant. Without providing parametric functional form assumptions, we provide assumptions on the manner in which errors depend on exogenous variables. In particular, as elaborated on below, we assume a generalized form of heteroscedasticity for both errors. We then develop a "feasible" control whose impact is not constant and show that the model is identified without exclusion restrictions.

As discussed in section 3, other papers have explored identification via second moments (e.g. Vella and Verbeek 1997, Rummery et al 1999, Sentana and Fiorentini 2001, Rigobon 2003 and Lewbel 2004). For the model that we consider, identification depends on there being heteroscedasticity in one or both equations of interest and that it "differs" across equations in a manner made precise below. The estimator is then based on estimating a generalized form of heteroscedasticity in each equation. For the structural equation of interest, such heteroscedasticity must be estimated simultaneously with the model's parameters as consistent residuals are unavailable. We do this in a setting where the conditional variance of each error is an unknown function of an index which needs to be estimated. While this semiparametric treatment of the unknown functions complicates the analysis, it ensures that we

[1] This control function approach is equivalent to two-stage-least-squares.

can consider a general form of heteroscedasticity without having to rely on parametric assumptions for identification.

In the following section we outline the model. In section 3 we discuss the estimation method and how to implement it. Formal results are stated in section 4. This section also outlines the proof strategy for obtaining these results. Section 5 provides simulation evidence and section 6 concludes. The Appendix contains detailed proofs of all theorems and intermediate lemmas.

2 Model and Identification Sources

With θ_o and π_o as vectors of true parameter values, consider the following linear triangular model:

$$\begin{aligned} Y_{1i} &= X_i\theta_{1o} + Y_{2i}\theta_{2o} + u_i \equiv W_i\theta_o + u_i \quad (1) \\ Y_{2i} &= X_i\pi_o + v_i, \quad (2) \end{aligned}$$

where Y_{1i} and Y_{2i} are continuous endogenous variables; X_i is a vector of variables that are mean-independent of the error components u_i and v_i. We further assume that these errors are correlated. The main objective of estimation is to conduct inference on θ_o, the vector of true parameter values in the primary equation. We use the terms primary and secondary to refer to the first and second equations respectively. Notice that the model allows the same $X's$ in both equations without imposing any restrictions on the parameter values.

When the errors do not depend on X, the (linear) relation between errors is captured by the following unconditional population regression:

$$a_o = \arg\min_a E\left[u - av\right]^2 \Rightarrow a_o = cov\left(u, v\right)/Var(v).$$

By construction, $\varepsilon \equiv u - a_o v$ is uncorrelated with v, which provides the basis for the controlled regression:

$$Y_{1i} = W_i\theta_o + a_o v_i + \varepsilon_i.$$

Provided that the matrix $[W,\ v\ ]$ has full column rank, the OLS estimator for this regression is consistent and would be implemented in practice by replacing v_i by the corresponding residual. However, in the absence of an exclusion restriction this full rank condition is not satisfied.

When the distribution of the errors depends on X, we would capture the (linear) conditional relation between the errors by the following conditional population regression:[2]

$$\begin{aligned} A_o(X) &= \arg\min_A E\left[u - Av \mid X\right]^2 \Rightarrow \\ A_o(X) &= cov\left(u, v \mid X\right)/Var(v \mid X). \end{aligned}$$

In this case, $\varepsilon \equiv u - A_o(X)\, v$ is uncorrelated with v conditioned on X, which provides the basis for the controlled regression:

$$Y_{1i} = W_i\theta_o + A_o(X_i)v_i + \varepsilon_i.$$

Provided that A_o depends on X, which would be reasonable when the error distributions depend on X, the matrix $[W \quad A_o(X)v]$ will have full column rank. Accordingly, when $A_o(X)$ is known, the above model is identified without exclusion restrictions.

As $A(X)$ is unknown, it must be estimated and restrictions must be imposed to obtain identification. Here, we explore the restrictions implied by a generalized form of heteroscedasticity. To this end, assume:

$$u \equiv S_u u^*; \quad v \equiv S_v v^*,$$

where

$$\begin{aligned} S_u^2 &\equiv Var\left(u|X\right) \\ S_v^2 &\equiv Var\left(v|X\right) \\ E\left(u|X\right) &= E(v|X) = 0. \end{aligned}$$

Further, there is a constant relation between unscaled error components:[3]

$$\rho_o \equiv E\left(u^*v^*|X\right) = E\left(u^*v^*\right).$$

Subject to the above restrictions, the error components can arbitrarily depend on X. With the correlation ρ_o constant, the control is given as:

$$A_o(X)v = \rho_o\left[S_{uo}/S_{vo}\right]v.$$

[2]We would like to thank Whitney Newey for this interpretation of the control.

[3]Note that Bollerslev (1990) also employs a constant correlation assumption in a time-series context.

Before discussing how to implement the above control, note that if the scaling functions are known or can be consistently estimated, then identification holds if these scaling functions "differ" in that the matrix $[S_{uo} \;\; S_{vo}]$ has full column rank. As a specialized interpretation, view u^* and v^* as unobserved variables with non-constant impacts on the endogenous variables. These impacts are functions of X and are given by the functions S_{uo} and S_{vo} respectively.

Other papers exploit second moment information as a source of identification. Vella and Verbeek (1997) and Rummery et al (1999) develop an estimation procedure based on the rank order of an individual's position in the reduced form residual distribution for subsets of the data. The variable determining the selection of subsets is also assumed to be responsible for the heteroscedasticity. In the context of normal factor models, Sentana and Fiorentini (2001) examine heteroscedasticity as a source of identification. Rigobon (2003) formulates a model in which there are two known regimes. The parameters of interest and the covariance between the equations' errors do not depend on the regime indicator. However, the error variances do depend on the known regime indicator. Employing an error covariance restriction similar to that in Rigobon, Lewbel (2004) examines a model of heteroscedasticity with second moment information depending on a known vector of variables Z. As Z may coincide with X, for comparative purposes we focus on this case and without loss of generality take $E(X) = 0$. He then considers a model in which:

$$E\left(X_i u_i v_i\right) = E\left[X_i E\left(u_i v_i | X_i\right)\right] = 0; \;\; E\left(X_i v_i^2\right) \neq 0.$$

The model considered here differs in several respects from those above. First, for the model outlined earlier, $E\left(u_i v_i | X_i\right)$ depends on X_i . Consequently, while the first restriction above may hold in special cases, it will not hold in general for the model considered here. Second, with the conditional covariance and variance functions depending on X_i, here the conditional variance of each error is modeled as an unknown function of an index. In a different model, Klein and Vella (2004) we also exploit heteroscedasticity to estimate a triangular treatment model where the endogenous regressor is binary. To flexibly model both the shape and conditional variance for the error distribution in the binary model, a double index formulation is employed. In so doing, with the estimated binary response probability as an instrument, the model is "well-identified" without exclusion restrictions.

While the treatment paper is related to this paper the identification and estimation strategies are fundamentally different from those employed here.

Finally note that the use of instruments in the absence of exclusion restrictions is not limited to cases of heteroscedasticity. Dagenais and Dagenais (1997) and Lewbel (1999) also discuss estimation of models where there are endogenous regressors and no exclusion restrictions. They show that when there is measurement error of a specific form one is able to use instruments based on the higher powers of the included variables. The model and estimator presented below both differ from the approach in these papers.

3 The Estimator: Implementing Strategies

Before presenting the main results, this section outlines and motivates the estimation strategy. From the above discussion, we will require residuals and the conditional variance function for the secondary equation. Accordingly, first obtain consistent estimates of the secondary equation's conditional mean parameter values by regressing Y_2 on X to get $\widehat{\pi}$. We then estimate the residuals as:[4]

$$\widehat{v} = Y_2 - X\widehat{\pi}.$$

To estimate S_v, we impose a single index structure:

$$S_{v_i}^2 \equiv E\left[v_i^2 \mid X_i\right] = E\left[v_i^2 \mid I_{vi}\left(\delta_o\right)\right],$$

where $I_{vi}\left(\delta_o\right) \equiv X_{1i} + X_{2i}\delta_o$. Next, estimate δ_o using semiparametric least squares with $\widehat{v}_i^2$ as the dependent variable (see Ichimura, 1993). Namely:

$$\hat{\delta} = \arg\min_{\delta} \sum \hat{\tau}_i \left[\hat{v}_i^2 - \hat{E}\left(\hat{v}_i^2 \mid I_{vi}\left(\delta\right)\right)\right]^2,$$

where $\hat{\tau}_i$ is a trimming functions that restricts X_i to a compact set.[5] Employing the estimated index:

$$\hat{S}_{vi}^2 = \widehat{E}\left(\widehat{v}_i{}^2 \mid I_{vi}\left(\hat{\delta}\right)\right)$$

where $\widehat{E}$ is a non-parametric estimator for the indicated conditional expectation. Employing the above initial estimator $\hat{S}_{vi}$, we then repeat the above

[4] These residuals could be obtained in a more general non-parametric or semiparametric regression.

[5] Here, the set will depend on sample quantiles for the $X\imath s$.

process in a GLS step.[6] For notational convenience below, denote the vector of parameter estimates as: $\hat{\eta} \equiv \left(\hat{\pi}' \ \hat{\delta}'\right)'$. As our focus will be on the primary equation, we will refer to these parameters as nuisance parameters.

For the primary equation of interest, we employ an estimator for S_u that is itself a function of unknown parameters. Let $I_{ui}(b_o) \equiv X_{1i} + X_{2i} b_o$ and assume the single index assumption:

$$E\left[u_i^2 \mid X_i\right] = E\left[u_i^2 \mid I_{ui}(b_o)\right].$$

As the above function can not be estimated directly, define the function:

$$\hat{S}_{ui}^2(\theta, b) \equiv E\left[(Y_1 - W_i\theta)^2 \mid I_{ui}(b)\right]$$

where the estimated expectation is obtained from a kernel-based nonparametric regression of $(Y_1 - W_i\theta)^2$ on $I_{ui}(b)$. With $\alpha \equiv (\theta, \rho, b)$ and $i = 1, ..., N$ observations, define:

$$\begin{aligned} \hat{A}_i(\alpha; \hat{\eta}) &\equiv \rho\left[\hat{S}_{ui}(\theta, b) / \hat{S}_{vi}\right] \\ \hat{M}_i(\alpha; \hat{\eta}) &\equiv W\theta + \hat{A}(\alpha)\hat{v} \\ \hat{S}(\alpha) &\equiv \frac{1}{N}\sum_i \hat{\tau}_i \left[Y_{1i} - \hat{M}_i(\alpha; \hat{\eta})\right]^2. \end{aligned}$$

The estimator for the primary equation is now defined as:

$$\hat{\alpha} \equiv \arg\min_{\alpha} \hat{S}(\alpha).$$

In the next section, we provide assumptions and definitions under which these estimators are consistent and asymptotically distributed as normal. With detailed proofs relegated to an appendix, this section also summarizes the main results and outlines the proof strategy for obtaining them.

4 Assumptions, Definitions, and Results

We employ the following assumptions to establish the asymptotic results proved in the Appendix:

[6] While it is possible to avoid a GLS step, we have found that the estimator for S_{vi} based on $\hat{\pi}_{GLS}$ is improved and that there is a corresponding improvement in the estimates of the primary equation of interest.

A1 The vector $(Y_{1i}, Y_{2i}, X_i, u_i, v_i)$ is i.i.d distributed over i, with $E(u_i^4|X_i)$ and $E(v_i^4|X_i)$ bounded.

A2 The parameter vector: $\gamma \equiv (\pi, \theta, \delta, b, \rho)$ is in a compact parameter space, Θ, where γ_o is in the interior of Θ.

A3 Write the error components as:

$$\begin{aligned} u &\equiv S_u u^*,\ S_u^2 \equiv Var(u|X) \\ v &\equiv S_v v^*,\ S_v^2 \equiv Var(v|X). \end{aligned}$$

Assume:

$$E(u^*|X) = E(v^*|X) = 0$$
$$\rho_o \equiv E(u^* v^*|X) = E(u^* v^*).$$

A4 With $I_{ui}(b_o) \equiv X_{1i} + X_{2i} b_o$ and $I_{vi}(\delta_o) \equiv X_{1i} + X_{2i}\delta_o$:

$$\begin{aligned} S_u^2 &\equiv E\left(u^2|X\right) = E\left[u^2 \mid I_{ui}(b_o)\right] > 0 \\ S_v^2 &\equiv E\left(v^2|X\right) = E\left[v^2 \mid I_{vi}(\delta_o)\right] > 0. \end{aligned}$$

For X in a compact set, these functions and their first four derivatives are uniformly bounded. Further, each index depends on a continuous variable.

A5 The conditional density $g(x_1|x_2)$ is bounded away from zero on the interior of its support and has bounded derivatives up to the fourth order.

A6 For estimating expectations and densities, assume that the kernel function, K, is a symmetric density with up to 4 bounded derivatives.[7]

A7 The matrix $[X,\ Y_2,\ (S_u/S_v)\, v] : N$ x $(K+2)$ has full column rank.

The above assumptions are somewhat standard, with the last assumption required for identification (see Theorem 2 of this section). In addition to these assumptions, we need to define the estimators and a bias reduction device used to establish asymptotic normality. Accordingly, we adopt the following definitions:

[7] In the simulations, K is a normal kernel.

D1 Indicator Trimming. Let $\underline{c}_k$ and $\bar{c}_k$ be lower and upper population quantiles for $X_{ik}, k = 1, ..., K$. Let q_o be the vector of these quantiles. With x: 1xK, define $\mathcal{P} \equiv \{x : \underline{c}_k < x_k < \bar{c}_k,\ k = 1, ..., K\}$. With $X_i \equiv [X_{i1}, ...X_{iK}]$, define the trimming indicator:

$$\tau_{io} \equiv \ \tau_i(q_o) \equiv \ \{ X_i \in \mathcal{P} \}.$$

With $\hat{q}$ as a vector of sample quantiles, the estimated trimming function is given as: $\hat{\tau}_i \equiv \tau_i(\hat{q})$.

To insure that various estimated denominators are bounded away from zero in large samples, the above trimming function restricts the components of X to a compact set depending on estimated sample quantiles. As a result, the trimming function should be viewed as being estimated rather than taken as fixed.[8] Employing this trimming function, (D2-4) provide the estimators for the parameter values of the secondary and primary equations.

D2 $\mathbf{Y_2}$-Model. Let $\hat{\pi}$ be the GLS estimator from the regression of Y_2 on X.[9] Define the residual:

$$\hat{v} \equiv Y_2 - X\hat{\pi}.$$

The estimated index parameters of the conditional error variance are then given by:

$$\hat{\delta} = \arg\min_{\delta} \hat{R}(\delta), \ \hat{R}(\delta) \equiv \sum_{i=1}^{N} \hat{\tau}_i \left[\hat{v}_i^2 - \hat{E}\left(\hat{v}_i^2 \mid I_{vi}(\delta)\right)\right]^2 / N,$$

where $\hat{E}$ is a nonparametric estimated expectation defined below. With $\hat{\eta} \equiv \left(\hat{\pi}, \hat{\delta}\right)$ estimating $\eta_o \equiv (\pi_o, \delta_o)$, we refer to $\hat{\eta}$ as the (nuisance) vector of estimates for the secondary equation.

---

[8] As discussed in Lemma G1 of the Appendix, we will be able to take estimated trimming as known under a result due to Pakes and Pollard (1989).

[9] First, obtain OLS residuals $\hat{v}_i$. Second, obtain $I_{vi}\left(\hat{\delta}\right)$, from the SLS estimator of δ_o. Next, define estimate $\hat{S}_{vi}^2$:

$$\hat{S}_{vi}^2 = \hat{E}\left(\hat{v}_i^2 \mid I_v\left(\hat{\delta}\right)\right).$$

Reweighting observations in the Y_2 model provides the GLS estimator of π_o. All of the results in this paper hold using the OLS estimator of π_o. We have found, however, that the finite sample properties of the estimator for the Y_1 model are improved in employing the GLS estimator for the secondary equation.

D3 Estimated Conditional Variance. With $\hat{\delta}$ given in (D2):

$$\hat{S}_{vi}^2 \equiv \hat{E}\left(\hat{v}_i^2 | I_{vi}\left(\hat{\delta}\right)\right).$$

D4 $\mathbf{Y_1}$-Model. With the Y_1- model given as:

$$Y_1 = [X \quad Y_2]\,\theta_o + u \equiv W\theta_o + u,$$

define $u(\theta) \equiv Y_1 - W\theta$ and $\hat{S}_{ui}^2(\theta, b) \equiv \hat{E}[u^2(\theta) \mid I_{ui}(b)]$. Then, with $\alpha \equiv (\theta, b, \rho)$ and

$$\hat{M}_i(\alpha) \equiv W_i\theta + \rho\left[\hat{S}_{ui}(\theta, b)/\hat{S}_{vi}^2\right]\hat{v}_i,$$

for the primary equation:[10]

$$\hat{\alpha} = \underset{\alpha}{\arg\min}\hat{S}(\alpha),\ \hat{S}(\alpha) \equiv \frac{1}{2}\sum \hat{\tau}_i\left[Y_{1i} - \hat{M}_i(\alpha)\right]^2/N.$$

Both secondary and primary equations depend on a nonparametric estimator of a conditional expectation. Definitions (D5-9) provide this estimator and the underlying windows upon which it depends.

D5 Estimated Conditional Expectations. Let $Z \equiv X_1 + X_2\gamma$ and define:

$$\hat{E}[Y \mid Z = z_i; \gamma] \equiv \hat{f}_i/\hat{g}_i,$$

where with the kernel function satisfying the assumptions in (A6):

$$\begin{aligned}
\hat{f}_i &\equiv \hat{f}\left(z_i; \hat{L}, \gamma\right) \equiv \sum_{j\neq i}\frac{Y_j\hat{L}_j}{(N-1)\,h}K\left(\left[\frac{z_i - z_j}{h}\hat{L}_j\right]\right) \\
\hat{g}_i &\equiv \hat{g}\left(z_i; \hat{\lambda}, \gamma\right) \equiv \sum_{j\neq i}\frac{\hat{\lambda}_j}{(N-1)\,h}K\left(\left[\frac{z_i - z_j}{h}\hat{\lambda}_j\right]\right).
\end{aligned}$$

Here, $\hat{L}_j$ and $\hat{\lambda}_j$ are estimated local smoothing parameters that are defined below and are employed for bias control. Note that the local smoothing parameters depend on a pilot window, h_p, and that the above estimators also depend on a global window, h. Restrictions on the windows h and h_p are given in (D10) after local smoothing parameters have been defined.

[10] The factor 1/2 avoids having to account for a factor of 2 in a number of gradient expressions examined in the Appendix.

With the exception of the estimated local smoothing parameters, (D5) is a standard nonparametric estimator for a conditional expectation. For reasons detailed in Lemmas 2-3 of the Appendix, such local smoothing serves as a device to reduce the bias in the components of this estimator. Definitions (D6-9) provide the components needed to define estimated local smoothing parameters.

D6 Smooth Trimming. Local smoothing parameters will be smoothly trimmed away from zero using the smoothed trimming function:

$$\tau_N(l;a) \equiv [1+\exp[\ (-Ln(N)Ln(N)\ )(l-a)\ ]\ ]^{-1}.$$

D7 Pilot Estimators. The estimated smoothing parameters are based pilot estimators. Define the following pilot estimators for the density $g(z)$ and the conditional expectation $E(Y|Z=z)$:

$$\begin{aligned}
\hat{g}_p(z_j) &\equiv \sum_{k\neq j}\frac{1}{(N-1)h_p}K\left(\left[\frac{z_j-z_k}{h_p}\right]\right)\\
\hat{E}_p(Y\mid Z=z_j) &\equiv \sum_{k\neq j}\frac{Y_j}{(N-1)h_p}K\left(\left[\frac{z_j-z_k}{h_p}\right]\right)/\hat{g}_p(z_j)\\
\hat{f}_p(z_j) &\equiv \hat{E}_p(Y|Z=z_j)\hat{g}_p(z_j),
\end{aligned}$$

where h_p is termed a pilot window and z_j is an index.

D8 Trimmed Pilot Estimators. Refer to the trimming function in (D1) and select population quantiles:

$$\underline{c}_k^* < \underline{c}_k,\ \bar{c}_k^* > \bar{c}_k$$

Define $\mathcal{P}^* \equiv \{x : \underline{c}_k^* < x_k < \bar{c}_k^*,\ k=1,...,K\}$ so as to contain $\mathcal{P}$ in (D1) as a proper subset. Then, define:

$$\underline{f} \equiv \inf_{\mathcal{P}^*}\hat{f}_p(z_j);\quad \underline{g} \equiv \inf_{\mathcal{P}^*}\hat{g}_p(z_j).$$

Referring to (D7), define the trimmed estimators

$$\begin{aligned}
\hat{f}_i^* &\equiv \left[1-\tau_N\left(\hat{f}_p(z_i);\underline{f}\right)\right]\underline{f}+\tau_N\left(\hat{f}_p(z_i);\underline{f}\right)\hat{f}_p(z_i)\\
\hat{g}_i^* &\equiv \left[1-\tau_N\left(\hat{g}_p(z_i);\underline{g}\right)\right]\underline{g}+\tau_N\left(\hat{g}_p(z_i);\underline{g}\right)\hat{g}_p(z_j).
\end{aligned}$$

D9 Estimated Local Smoothing Parameters. Letting m_1 and m_2 be the geometric means of $\hat{f}_j^*$ and $\hat{g}^*(z_j)$ respectively, the local smoothing parameters in (D5) are given by:

$$\hat{L}_j \equiv \left[\hat{f}_j^j/m_1\right]^{1/2}; \quad \hat{\lambda}_j \equiv \left[\hat{g}_j^*/m_2\right]^{1/2}.$$

If the local smoothing parameters were known and bounded away from zero, then from Lemma 2-3 in the Appendix they would serve as an appropriate bias reducing device. Trimming is required to insure that these parameters are bounded away from zero. In this context, smooth rather than indicator trimming is required for technical reasons. With a smooth trimming function approximating an indicator, the derivative of this function must increase without bound as the sample size increases. The $Ln(N)^2$ factor of the trimming function insures that this derivative slowly increases.

D10 Windows. Set pilot and second stage global windows such that:

$$h_p \equiv \hat{\sigma}N^{-r_p}; \quad h \equiv \hat{\sigma}N^{-r},$$

$\hat{\sigma} = O_p\left(N^{-1/2}\right)$.[11] Assume:

$$1/8 < r < 1/6; \quad r > r_p; \quad r + r_p > 1/2.$$

The main asymptotic results of this paper follow from the above assumptions and definitions and are proved in the Main section of the Appendix. In the remainder of this section, we summarize the asymptotic results and provide a brief outline of the proof strategy. Detailed arguments are given in the Appendix. Beginning with the secondary equation (Y_2-Model), Theorem 1 provides the large sample results for the estimators of the nuisance parameters.

Theorem 1 (The Y_2-Model). Under the above assumptions and definitions, estimates of regression and index parameters satisfy the characterizations:

$$1) \ \sqrt{N}[\hat{\pi} - \pi_o] = \sqrt{N}\sum_{i=1}^{N} \varepsilon_{\pi i}/N$$

$$2) \ \sqrt{N}\left[\hat{\delta} - \delta_o\right] = \sqrt{N}\sum_{i=1}^{N} \varepsilon_{\delta i}/N,$$

[11] The estimated component $\hat{\sigma}$ need not be $\sqrt{N}-$convergent, and may differ for pilot and second stage windows.

where $\varepsilon_{\pi i}$ and $\varepsilon_{\delta i}$ each are $i.i.d.$ with 0 expectation and finite variance.

The first result above is immediate and the second follows from a standard Taylor series argument and Ichimura (1993). This result also follows from the same type of U-statistic arguments used to establish asymptotic normality for estimates of the primary equation (Y_1-Model).

For the Y_1-Model, Theorem 2 below provides the consistency/identification result.

Theorem 2 (Consistency: the $\mathbf{Y_1}-$Model). With $\alpha_o \equiv (\theta_o, \rho_o, b_o)$ and $\hat{\alpha} \equiv \left(\hat{\theta}, \hat{\rho}, \hat{b}\right)$, under the above assumptions and definitions:

$$\hat{\alpha} \xrightarrow{p} \alpha_o.$$

To outline the consistency argument, which is provided in detail in the Appendix, recall from D4 that:

$$\hat{\alpha} = \arg\min_{\alpha} \hat{S}(\alpha).$$

Equivalently, employing a well-known device that avoids having to establish convergence for $\hat{S}(\alpha_o)$:

$$\hat{\alpha} = \arg\min \hat{Q}(\alpha), \ \hat{Q}(\alpha) \equiv \hat{S}(\alpha) - \hat{S}(\alpha_o)$$

Referring to (D4), obtain $M_i(\alpha)$ from $\hat{M}_i(\alpha)$ by replacing all estimated functions with their uniform probability limits. Then, define $Q(\alpha)$ by replacing $\hat{M}_i(\alpha)$ in $\hat{Q}(\alpha)$ with $M_i(\alpha)$. It can be shown that

$$\left|\hat{Q}(\alpha) - Q(\alpha)\right| \text{ and } \left|\hat{Q}(\alpha) - E[Q(\alpha)]\right|$$

each converge in probability, uniformly in α to zero. Accordingly, consistency will follow if $E[Q(\alpha)]$ is uniquely minimized at α_o. With $M_{oi} \equiv M_i(\alpha_o)$, $\Delta_i(\alpha, \alpha_o) \equiv 2(M_i - M_{io})$, and with $\Delta(\alpha, \alpha_o)$ as the corresponding vector of differences:

$$E[Q(\alpha)] = E[\Delta(\alpha, \alpha_o)]'[\Delta(\alpha, \alpha_o)]/N.$$

As $\Delta(\alpha, \alpha_o) = 0$ at $\alpha = \alpha_o$, α_o is a minimizer and the issue is one of uniqueness. Under a constant correlation assumption, in the appendix we

establish identification (uniqueness) when the matrix $[X, Y_2, (S_u/S_v)v]$ has full column rank.

Theorem 3 below, which is proved in the appendix, provides the normality result.

Theorem 3 (Normality: the $\mathbf{Y_1}-$Model). Under the assumptions and definitions above:

$$\sqrt{N}\left[\hat{\alpha}-\alpha_o\right] \xrightarrow{d} Z,\ Z \tilde{} N\left(0,\Sigma\right).$$

To outline the argument, note that under a standard Taylor series argument for the gradient to the objective function and a uniform convergence argument for the Hessian, normality will follow if the normalized gradient is asymptotically distributed as normal. For expositional purposes, neglect first-stage estimation uncertainty as it matters,[12] but poses no technical difficulties. Then, evaluating all functions at true parameter values and letting $\hat{w}_i \equiv \nabla_\alpha \hat{M}_{oi}$ the gradient is given as:

$$\begin{aligned}\sqrt{N}\hat{G} &= -\sqrt{N}\sum \hat{\tau}_i\left[Y_{1i}-M_{oi}\right]\hat{w}_i/N + \sqrt{N}\sum \hat{\tau}_i\left[\hat{M}_{oi}-M_{oi}\right]\hat{w}_i/N \\ &\equiv \sqrt{N}\hat{G}_1\left(\alpha_o\right)+\sqrt{N}\hat{G}_2\left(\alpha_o\right).\end{aligned}$$

With the estimated trimming function depending on sample quantiles, Lemma G1 employs results in Pakes and Pollard (1989) to show that the trimming function may be taken as known. A mean-square convergence argument is then used to show that the estimated weight function, $\hat{w}_i$, may be taken as known. Accordingly: with $\varepsilon_{1i} \equiv \left[Y_{1i}-M_{oi}\right]w_i$ and with $\bar{\varepsilon}_1$ as the corresponding sample mean:

$$\sqrt{N}\hat{G}_1 = \sqrt{N}\bar{\varepsilon}_1 + o_p(1).$$

For the second gradient component, Lemma 5 shows that the estimated trimming and weight functions may be taken as known:

$$\sqrt{N}\hat{G}_2 = \sqrt{N}\sum \tau_{io}\left[\hat{M}_{oi}-M_{oi}\right]w_i/N + o_p(1).$$

[12] See Newey and McFadden (1994).

With $\hat{M}_{oi}$ depending on estimated expectations, the bias properties of these expectations are important to the argument. Recall that an estimated expectation, which is calculated under local smoothing, has the form:

$$\hat{E}_i \equiv \hat{f}_i/\hat{g}_i.$$

Many of the intermediate lemmas in the Appendix are concerned with showing that the bias in numerator and denominator is sufficiently small. Once it is established that local smoothing delivers a bias of order $o(N^{-1/2})$, Lemmas 7 and G2 of the Appendix establish that $\sqrt{N}\hat{G}_2$ is close in probability to a linear combination of U-statistics. From a standard projection argument, it is then possible to characterize this gradient component in the same form as the first component.[13] Namely in the Appendix we define a vector ε_{2i} which is i.i.d. with expectation zero and finite variance components. Then, with $\bar{\varepsilon}_2$ as the corresponding sample mean, we show that:

$$\sqrt{N}\hat{G}_2 = \sqrt{N}\bar{\varepsilon}_2 + o_p(1).$$

With the estimation uncertainty component having a similar i.i.d. mean characterization, asymptotic normality follows.

5 Simulation Evidence

To analyze the performance of the estimator we examine the following setting. We simulated the following model where the same exogenous variables appear in the conditional means and the conditional variances of both endogenous variables. The two indices underlying the heteroscedasticity are also highly correlated. Moreover, we use the same functional form for the heteroscedasticity in each equation. The model has the form:

$$\begin{aligned}
Y_{1i} &= 1 + x_{1i} + x_{2i} + Y_{2i} + u_i \\
Y_{2i} &= 1 + x_{1i} + x_{2i} + v_i \\
u_i &= 1 + \exp(.2 * x_{1i} + .6 * x_{2i}) * u_i^* \\
v_i &= 1 + \exp(.6 * x_{1i} + .2 * x_{2i}) * v_i^* \\
u^* &= .33 * v_i^* + N(0,1) \text{ and } v_i^* \sim N(0,1).
\end{aligned}$$

[13] See Serfling (1980) and Powell, Stock, and Stoker (1989).

We examine two distributions for the $x_i's$. In the first design we generate both x_{1i} and x_{2i} as standard normal random variables. In the second design we retain x_{1i} but transform x_{2i} into a chi-squared variable with 1 degree of freedom. We then estimated the model by OLS and the control function procedure developed here, which we denote CF in the tables. The simulation results for $n = 1000$ and 100 replications are reported in Table 1.

An examination of Table 1 reveals a number of interesting features of the simulations. First, turn to the left hand side of Table 1 which reports results for the normal design. The entries in the Table represent the mean value from the 100 replications with the standard deviation of the replications reported in parentheses under the estimate. The OLS estimates for the main equation's parameters in this specification are badly biased with respect to their true values of 1 indicating that there is a large degree of endogeneity in this model. The estimates for each of the $x's$ are approximately .7, indicating a bias of around 30 percent, while the bias on the coefficient for the endogenous regressor is approximately 29 percent.

The lower panel of columns 1 shows results for the control function procedure under the normal design. For the δ-parameter underlying the index generating heteroscedasticity in the secondary equation, the mean of the estimates is .351. This value is reasonably close to the true value of .33, but there is a relatively large standard deviation. The parameter b is the coefficient in the index generating heteroscedasticity in the main equation. The average point estimate of .261 is reasonable relative to the true value of .33, but note that there is a very large standard deviation associated with this estimate. Recall that this estimate is obtained simultaneously with the slope coefficients and its imprecision reflects that it is difficult to estimate this parameter accurately while simply minimizing the squared residuals for this model. If this parameter is of direct interest, then as described below it is possible to exploit other sources of information to increase its precision. The parameter ρ corresponds to the coefficient on the control function. Given the design the true value is .33 and thus the average estimate of .304 is good. Most importantly, however, is that the average value of the coefficients for the $x's$ and Y_2 are all close to 1 indicating that the inclusion of the control function is accounting for the endogeneity bias. Moreover, while there is more variability in the estimates, in comparison to the OLS estimates, the estimates are generally quite accurate as indicated by their standard deviations.

In the remainder of the Table we report the corresponding estimates for

the chi-square design. The OLS estimates indicate that with the chi-squared x the problem of endogeneity is reduced. The remaining estimates are generally consistent with the normal design.

Though not reported here, we also considered several variants on the estimator presented. Here, we have explored the case in which the conditional variance of the errors is characterized by a single index. However, suppose that the entire distribution of each error is characterized by a single index (as is the case in the simulations). Under this more restrictive index assumption, it is possible to develop a modified version of the estimator presented here with significantly better finites sample performance (especially for index parameters).[14] We have also examined several "GLS" variants of the CF method presented here. While these resulted in a noticeable improvement in the estimates, we judged the improvement not sufficient to warrant any further (albeit minor) lengthening of the Appendix.[15]

[14] When the entire distribution of the errors depends on a single index, any function of the squared residual will satisfy a single index assumption. Accordingly, in an SLS regression, there will be many ways of estimating the index parameters. For the multiplicative heteroscedasticity employed in the monte-carlo, the (trimmed) log transform of the squared residual would appear to a natural transformation to employ, and we found that it resulted in a noticeable finite sample improvement estimates of both secondary and primary equations.

[15] We examined estimators based on consistent residuals from the primary equation to estimate the conditional variance parameter in a manner similar to that employed for the secondary equation. Not surprisingly, as the information on the conditional variance is largely contained in the residuals, the resulting estimator for the index parameter had a much smaller variance that obtained from the control method reported here. Indeed, it would be interesting to combine this moment information with that in the first-order conditions to the minimization problem defined here. We have not explored this possibility in the present paper.

Table 1: Simulation Results

	Normal	**Chi-sq**
	OLS	**OLS**
constant	.704	.858
	(.058)	(.122)
x_1	.706	.858
	(.049)	(.120)
x_2	.710	.866
	(.060)	(.121)
Y_2	1.291	1.137
	(.039)	(.108)
	CF	**CF**
constant	1.023	1.032
	(.278)	(.304)
x_1	1.026	1.034
	(.276)	(.302)
x_2	1.032	1.042
	(.281)	(.298)
Y_2	.975	.963
	(.273)	(.305)
ρ	.304	.302
	(.184)	(.200)
b	.261	.406
	(.717)	(.848)
δ	.351	.361
	(.158)	(.242)

6 Conclusion

In summary, we have examined a generalized form of heteroscedasticity with conditional covariance and variance functions depending on exogenous variables. For this case, we have shown that the model is identified and have formulated a method for estimating it. We have established that the estimator is consistent and asymptotically distributed as normal. In a monte-carlo study, across several designs, the estimator for the parameters of interest in the primary equation performed quite well in finite samples. As indicated previously, there is scope for further improvements in the estimator for the index parameters. Such improvements would come from fully exploring index structure in the monte-carlo (see footnote 11) or from making use of all available moment information (see footnote 12).

We have focused on the linear structure in part because it is most often used in practice. More importantly, in the absence of other information, it is this structure for which identification fails without exclusion restrictions. Nevertheless, it would seem relatively straight-forward to extend the model to allow nonlinear functions of the exogenous variables to enter both primary and secondary equations. With a modified control based on $E(u|v, X)$, it would also seem possible to allow for nonlinearities in the endogenous variables.

7 Appendix

7.1 Intermediate Lemmas

The Appendix is organized into a section on intermediate lemmas and a main section providing consistency and asymptotic normality for the proposed estimator. In this section of the Appendix, we provide a characterization for the gradient from which asymptotic normality will follow. To preview the argument, we require the following notation. Denote $\hat{\eta}$ as the vector of estimated parameters from the Y_2-equation (including estimated parameters of the conditional error variance). Write $\alpha_o \equiv (\theta_o', \rho_o, b_o')'$ for the column vector of true parameter values from the Y_1-model (including correlation and conditional variance parameter values). Refer to the averaged objective function for the primary equation (Y_1-model) shown in (D4). Taken with respect to α let $\hat{G}(\alpha_o; \hat{\eta})$ and $\hat{H}(\alpha_o; \hat{\eta})$ be the corresponding gradient and Hessian.

Employing the above notation, from a Taylor series expansion:

$$\sqrt{N}[\hat{\alpha} - \alpha_o] = -\hat{H}\left(\alpha^+; \hat{\eta}\right)^{-1} \sqrt{N}\hat{G}(\alpha_o; \hat{\eta}), \tag{1A}$$

$\alpha^+ \epsilon [\alpha_o . \hat{\alpha}]$. Obtain $H(\alpha; \eta)$ from $\hat{H}$ by replacing all estimated nonparametric expectations by their probability limits in $\hat{H}$. Uniformly in the parameters, it can be shown that $\left|\hat{H}(\alpha; \eta) - H(\alpha; \eta)\right|$ and $|H(\alpha; \eta) - EH(\alpha; \eta)|$ each converge in probability to zero. Therefore, once consistency is established, $\hat{H}(\alpha^+; \hat{\eta})$ will converge in probability to $H_o \equiv EH(\alpha_o; \eta_o)$. Asymptotic normality will then follow if the gradient component is asymptotically normal.

To analyze the gradient, with $\hat{M}_i(\alpha_o, \hat{\eta}) \equiv W_i\theta_o + \hat{A}_i(\alpha_o; \hat{\eta})\hat{v}_i$, define:

$$\begin{aligned} \hat{M}_{oi} &\equiv \hat{M}_i(\alpha_o, \eta_o) \equiv W_i\theta_o + \hat{A}_i(\alpha_o; \eta_o)v_i \\ M_{oi} &\equiv W_i\theta_o + A_i(\alpha_o; \eta_o)v_i. \end{aligned}$$

Denote $\hat{w}_i \equiv \nabla_\alpha \hat{M}_i(\alpha_o, \eta_o)$ and $\hat{G}_3 \equiv \left[\nabla \hat{G}_\eta(\theta_o, b_o; \eta^+)\right][\hat{\eta} - \eta_o]$. Then, from (D4) , the gradient with respect to α at α_o is given as:

$$\hat{G}(\alpha_o; \hat{\eta}) = \hat{G}(\alpha_o; \eta_o) + \hat{G}_3 \tag{2A}$$

$$
\begin{aligned}
&= -\sum \hat{\tau}_i \left[Y_{1i} - M_{oi}\right] \hat{w}_i/N + \sum \hat{\tau}_i \left[\hat{M}_{oi} - M_{oi}\right] \hat{w}_i/N + \hat{G}_3 \\
&\equiv \hat{G}_1 + \hat{G}_2 + \hat{G}_3.
\end{aligned}
$$

Lemmas G1 and G2 below provide appropriate characterizations for $\sqrt{N}\hat{G}_1$ and $\sqrt{N}\hat{G}_2$. The required characterization for the third component will immediately follow from Theorem 1 in the next section which characterizes the first-stage estimator, $\hat{\eta}$. The other intermediate lemmas in this section provide results required for these characterizations.

For a number of terms, we will need to deal with the fact that the trimming function depends on estimated sample quantiles. For one such term, we provide a complete argument in Lemma G1 to show that the trimming function may be taken as known. As a similar argument applies to other terms, subsequently we will take the trimming function as known.

In what follows, we will make use of several standard convergence results. First, recall from (D5) that with $Z \equiv X_1 + X_2\gamma$, an estimated conditional expectation has the form:

$$
\hat{E} \equiv \hat{E}\left[Y \mid Z = z; \gamma\right] \equiv \hat{f}/\hat{g}.
$$

Defining the derivative operator on a function f as:

$$
\nabla_\gamma^d (f) \equiv \frac{\partial}{\partial \gamma} f, \ \nabla_\theta^0 f \equiv f,
$$

it can be shown that

$$
\sup_{z,\,\gamma} \left| \nabla_\gamma^d \hat{E} - \nabla_\gamma^d E \right| \xrightarrow{p} 0, \ d = 0, 1, 2.
$$

With $E \equiv f/g$, estimated local smoothing functions depend on $\hat{f}_p$ and $\hat{g}_p$, pilot estimates of f and g respectively (D8-9). Obtain $\hat{f}^*$ from $\hat{f}$ by evaluating local smoothing functions at the expectation of pilot estimators. Then, for $d = 0, 1, 2$ and with all functions evaluated at γ_o :

$$
\begin{aligned}
\sum \left[\nabla_\gamma^d \hat{f}_i - \nabla_\gamma^d \hat{f}_i^*\right]^2 /N &= O_p\left(r_1\right), \ r_1 = \frac{1}{N h_p^d} \\
\sum \left[\nabla_\gamma^d \hat{f}_i^* - E\left(\nabla_\gamma^d \hat{f}_i^*\right)\right]^2 /N &= O_p\left(r_2\right), \ r_2 = \frac{1}{N h^d} \\
\sum \left[E\left(\nabla_\gamma^d \hat{f}_i^*\right) - \nabla_\gamma^d f_i\right]^2 /N &= O\left(r_3\right), \ r_3 = h^4.
\end{aligned}
$$

The first rate above follows from a Taylor series expansion, Lemma 1 below, and the rate at which the variance of a pilot estimator tends to zero. The other rates follow from the rates at which the expectation of each sum tends to zero. It follows that:[16]

$$\sum \left[\nabla_\theta^d \hat{f}_i - \nabla_{\theta_i}^d f_i\right]^2 / N = O_p\left(\max\left(r_i\right)\right),\ i = 1, 2, 3.$$

Similar results hold for the denominator for the estimated density $\hat{g}_i$. Lemma 1 below provides a basic result used in analyzing gradient components based on the above rates.

Lemma 1. Assume:

$$S_a \equiv \sum \hat{a}_i^2 / N = O_p\left(N^{-s}\right),\ S_b \equiv \sum \hat{b}_i^2 / N = O_p\left(N^{-t}\right),$$

where $s + t > 1$. Then,

$$\sqrt{N} \sum \hat{a}_i \hat{b}_i = o_p(1).$$

Proof of Lemma 1. The result follows from Cauchy's inequality:

$$\left|\sqrt{N} \sum \hat{a}_i \hat{b}_i / N\right| \leq \sqrt{N} S_a^{1/2} S_b^{1/2}.$$

Lemma G1 (First Gradient Component). With $W_i \equiv (X_i,\ Y_{2i})$ and $Y_{1i} - M_{oi} = u_i - \rho_o \left(S_{ui}/S_{vi}\right) v_i$:

$$\hat{w}_i \equiv - \begin{bmatrix} W_i + \rho_o \left(\nabla_\theta \hat{S}_{ui}\right) v_i / \hat{S}_{vi} \\ \left(\hat{S}_{ui}/\hat{S}_{vi}\right) v_i \\ \rho_o \left(\nabla_b \hat{S}_{ui}\right) v_i / \hat{S}_{vi}. \end{bmatrix}.$$

[16] With $\Delta \equiv \left[\nabla_\theta^d \hat{f}_i - \nabla_\theta^d \hat{f}_i\right]$, let:

$$\Delta_1 \equiv \left[\nabla_\theta^d \hat{f}_i - \nabla_\theta^d \hat{f}_i^*\right];\ \Delta_2 \equiv \left[\nabla_\theta^d \hat{f}_i^* - \nabla_\theta^d f_i\right].$$

Since $\Delta_1^2 + \Delta_2^2 > 2\Delta_1\Delta_2,\ \Delta^2 \leq 2\left(\Delta_1^2 + \Delta_2^2\right)$. The rate follows by writing Δ_2 as a sum of appropriate terms.

Define w_i by replacing $\hat{S}_{ui}$ with S_{ui} and $\hat{S}_{vi}$ with S_{vi} in $\hat{w}_i$. With $\tau_{io} \equiv \tau_i(q_o)$, let:

$$\varepsilon_{1i} \equiv -\tau_{io}[Y_{1i} - M_{oi}]\, w_i;\ \bar{\varepsilon}_1 \equiv \sum_{i=1}^{N} \varepsilon_{1i}/N.$$

Then, $\sqrt{N}\hat{G}_1 \equiv -N^{-1/2}\sum \hat{\tau}_i [Y_{1i} - M_{oi}]\, \hat{w}_i$:

$$\sqrt{N}\hat{G}_1 \equiv -N^{-1/2}\sum \hat{\tau}_i [Y_{1i} - M_{oi}]\, \hat{w} = \sqrt{N}\bar{\varepsilon}_1 + o_p(1).$$

Proof of Lemma G1. With $\hat{\tau}_i \equiv \tau_i(\hat{q})$, $\sqrt{N}\hat{G}_1$ is the sum of the following three terms:

$$\begin{aligned}
\mathbf{A} &\equiv N^{-1/2}\sum [Y_{1i} - M_{oi}]\, [\hat{\tau}_i - \tau_{io}]\, w_i \\
\mathbf{B} &\equiv N^{-1/2}\sum [Y_{1i} - M_{oi}]\, [\hat{\tau}_i - \tau_{io}]\, [\hat{w}_i - w_i] \\
\mathbf{C} &\equiv N^{-1/2}\sum [Y_{1i} - M_{oi}]\, \tau_{io}\, [\hat{w}_i - w_i].
\end{aligned}$$

The proof will follow if each of these terms is $o_p(1)$. Employing the same strategy as in Klein (1993), denote q_o as a vector of population quantiles (see (D1), Section 4) and let $N_\varepsilon \equiv \langle q : |q - q_o| < \varepsilon \rangle$, $\varepsilon = o(1)$. Then, $\mathbf{A} = o_p(1)$ if

$$\mathbf{A}^* \equiv \sup_{N_\varepsilon} N^{-1/2}\sum [Y_{1i} - M_{oi}]\, [\tau_i(q) - \tau_i(q_o)]\, w_i = o_p(1)$$

for all $\varepsilon = o(1)$.[17] From Pakes and Pollard [1989, Lemma 2.17, p. 1037], $\mathbf{A}^* = o_p(1)$.

For the term $\mathbf{B}$, it suffices to show that:

$$\mathbf{B}^* \equiv \sup_{N_\varepsilon} N^{-1/2}\sum [Y_{1i} - M_{oi}]\, [\tau_i(q) - \tau_i(q_o)]\, [\hat{w}_i - w_i] = o_p(1).$$

Letting

$$\tau_i^*(q) = \begin{cases} 1: & \tau_i(q) = 1 \text{ and/or } \tau_i(q_o) = 1 \\ 0: & Otherwise, \end{cases},$$

[17] If uniformity holds for $\alpha \in \mathcal{N}_\varepsilon$ for all $\varepsilon = o(1)$, then uniformity holds over $o_p(1)$ neighborhoods of q_o.

it follows from the definition of $\tau_i^*(q)$ and Lemma 1 that:

$$\begin{aligned}
\mathbf{B}^* &\leq N^{1/2}\mathbf{B}_1^*\mathbf{B}_2^*, \\
\mathbf{B}_1^* &= \sup_{N_\varepsilon}\left[\sum \left[Y_{1i} - M_{oi}\right]^2 \left[\tau_i(q) - \tau_i(q_o)\right]^2 / N\right]^{1/2} \\
\mathbf{B}_2^* &= \sup_{N_\varepsilon}\left[\sum \tau_i^*(q)\left[\hat{w}_i - w_i\right]^2 / N\right]^{1/2}.
\end{aligned}$$

From Klein (1993), with indicators approximated by smooth functions, it can be shown that for any fixed δ arbitrarily small : $\mathbf{B}_1^* = o_p\left(N^{-1/2+\delta}\right)$. It can be shown that $\mathbf{B}_2^* = o_p\left(N^{-\delta}\right)$, which completes the argument for $\mathbf{B}$.

Turning to $\mathbf{C}$, the analysis is similar to that in Klein and Spady (1993). For this term, it can be shown that $\hat{w}_i - w_i$ may be replaced by a linear combination of estimated functions. The result then follows from a mean-square convergence argument. To illustrate the argument, with $\varepsilon_i \equiv \left[Y_{1i} - M_{oi}\right] v_i$, one of the components of $\mathbf{C}$ is given as:

$$\mathbf{D}_1 \equiv N^{-1/2}\sum \tau_{io}\varepsilon_i\left(\hat{S}_{ui} - S_{ui}\right)/\hat{S}_{vi}.$$

From a Taylor series expansion and Lemma 1:

$$\mathbf{D}_1 = N^{-1/2}\sum \tau_{io}\varepsilon_i\left[\hat{f}_i/\hat{g}_i - f_i/g_i\right]/a_i + o_p(1), \; a_i \equiv S_{ui}S_{vi}.$$

It can now be shown that

$$\begin{aligned}
\mathbf{D}_1 &= \mathbf{D}_1^* + o_p(1), \\
\mathbf{D}_1^* &\equiv N^{-1/2}\sum \tau_{io}\varepsilon_i\left[\hat{f}_i/\hat{g}_i - f_i/g_i\right]\left[\hat{g}_i/g_i\right]/a_i.
\end{aligned}$$

Finally, it can be shown that $E\left[\left(\mathbf{D}_1^*\right)^2\right] \to 0$. A similar argument applies to the other terms of C, which completes the argument.

Abramson (1982) and Silverman (1986) provide a uniform bias result density estimation under known local smoothing. Hall (1998) extends this argument in several directions and provides a method for making explicit bias calculations in a variety of contexts. The Lemma below examines the extension for integrating with respect to a function that need not be a density.

Lemma 2 . Let $p(z)$ be a non-negative function with uniformly bounded derivatives to order 4 and let K be a symmetric kernel with a bounded derivatives to order 4. For $t \epsilon \mathcal{P}^*$, assume that $p(t) > 0$. Denote $\underline{p} \equiv \inf p(t)$, t $\epsilon \mathcal{P}^*$. Employing the smooth trimming function in (D6), define:

$$p^*(z) \equiv \tau\left(p(z); \underline{p}\right) p(z) + \left[1 - \tau\left(p(z); p\right)\right] \underline{p}.$$

Then, for $\mathcal{P}$ a proper subset of $\mathcal{P}^*$, uniformly in $t \epsilon \mathcal{P}$:

$$B\left(\hat{p}(t); p^*\right) \equiv \int \frac{p^*(z)^{1/2}}{h} K\left(\left(t - z\right)\frac{p^*(z)^{1/2}}{h}\right) p(z)\, dz - p(t) = O\left(h^4\right).$$

Proof of Lemma 2 . Make a change of variable with $\varepsilon \equiv (z - t)/h)$ and then take a Taylor series expansion about $h = 0$ to obtain:

$$B - p(t) = \sum_{i=1}^{3} h^i C_i^*(t) + O\left(h^4\right).$$

From symmetry of the kernel, terms in odd powers of h vanish. Since $\tau - 1$ and its derivatives vanish exponentially on $\mathcal{P}$, the second order term, $C_2^*(t)$, is arbitrarily close to $C_2(t)$, the second order term in the analogous expansion of $B\left[\hat{g}(t); g\right]$. For the case in which p is a density, Abramson and Silverman show that $\mathbf{C}_2(t) = 0$. As shown by Hall, this result also holds when p is not a density.

The estimator for $E(Y \mid Z = t)$ has the form $\hat{f}(t)/\hat{g}(t)$, with Lemma 2 providing the bias for $\hat{g}(t)$. For $\hat{f}(t)$, under known local smoothing with $L_j \equiv f_p^*(z_j)^{1/2}$ (see D8-9):

$$\hat{f}(t; L) \equiv \frac{1}{N-1} \sum_{j \neq i} y_j \left[\frac{L_j}{h} K\left(\frac{t - z_j}{h} L_j\right)\right].$$

Lemma 3 below, which is due to Hall (1990), examines the bias in $\hat{f}(t; L)$ with known local smoothing.

Lemma 3. Define $\hat{f}(t; L)$ and $f(t)$ as above as above. Assume that $f(t)$ has uniformly bounded derivatives up to order 4, and that the kernel

function, K, also has uniformly bounded fourth derivative. Assume that $f^*(t) > 0$ for $t \epsilon \mathcal{P}^*$. Let $\mathcal{P}$ be a proper subset of $\mathcal{P}^*$. Then, uniformly in $t \epsilon \mathcal{P}^*$:

$$E\left(\hat{f}(t;L)\right) = E(Y|Z=t)g_v(t) + O\left(h^4\right).$$

Proof of Lemma 3. Hall has proved this result in a much more general context. For this particular function and kernel employed here, the result follows from Lemma 2 with $p(z) \equiv [EY(|Z=z)g(z)]$.

Recall from (D7-9) that local smoothing functions depend on the pilot estimators $\hat{f}_p^*$ and $\hat{g}_p^*$. To deal with estimated local smoothing parameters we pursue the following strategy. First, we show in Lemma 4 that with local smoothing functions evaluated at their conditional expectations, then the low order bias results of Lemmas 2-3 still hold. Subsequently, we will show in Lemmas 5-6 that all gradient terms based on estimated local smoothing parameters are asymptotically close to those in which local smoothing functions are evaluated at the expectation of the pilot estimators. Define "fixed" local smoothing parameters:

$$\bar{L}_j \equiv \left[\bar{f}_j^*\right]^{1/2};\ \bar{\lambda}_j \equiv \left[\bar{g}_j^*\right]^{1/2}.$$

Here, with

$$\bar{f}_j \equiv E\left[\hat{f}_j|Y_i, X_i\right] \text{ and } \bar{g}_j \equiv E\left[\hat{g}_j|Y_i, X_i\right],$$

$\bar{f}_j^*$ and $\bar{g}_j^*$ denote smoothly trimmed versions of these densities as given in (D8).

An infeasible estimator for $E(Y \mid z_r)$ is given by $\bar{E}_r \equiv \hat{f}\left(z_r;\bar{L}\right)/\hat{g}\left(z_r;\bar{\lambda}\right)$:

$$\begin{aligned}
\hat{f}\left(z_r;\bar{L}\right) &= \sum_{j\neq i}^{N} Y_j K\left[\left(\frac{z_r - z_j}{h}\right)\bar{L}_j\right]\frac{\bar{L}_j}{(N-1)h}; \\
\hat{g}\left(z_r;\bar{\lambda}\right) &= \sum_{j\neq i}^{N} K\left[\left(\frac{z_r - z_j}{h}\right)\bar{\lambda}_j\right]\frac{\bar{\lambda}_j}{(N-1)h}.
\end{aligned}$$

Below, Lemma 4 provides an appropriately low order of the bias for the infeasible estimators $\bar{f}_r$ and $\bar{g}_r$.

Lemma 4. Let the pilot window for estimating local smoothing parameters be given as: $h_p = O(N^{-r_p})$ and the window for estimating second stage densities as: $h = O\left(N^{-r}\right)$, $r_p + r > 1/2$. Then for $\mathcal{P}$ and $\mathcal{P}^*$ as given above:

$$\begin{aligned} \sup_{t \,\epsilon\, \mathcal{P}^*} \Delta_1(t) &\equiv \sup \left| E\hat{f}\left(t; \bar{L}\right) - E\left[Y|Z=t\right] g(t) \right| = o\left(N^{-1/2}\right) \\ \sup_{t \,\epsilon\, \mathcal{P}} \Delta_2(t) &\equiv \sup \left| E\hat{g}\left(t; \bar{\lambda}\right) - g(t) \right| = o\left(N^{-1/2}\right). \end{aligned}$$

Proof of Lemma 4. Referring to Lemma 4, from a Taylor series expansion in h about $h = 0$:

$$\Delta_2(t) = \sum_{i=1}^{4} h^i \hat{C}_i(t).$$

From symmetry of the kernel, $\hat{C}_1 = \hat{C}_3 = 0$. For the second order term:

$$h^2 \hat{C}_2(t) = h^2 C_2(t) + h^2 \left[\hat{C}_2(t) - C_2(t)\right],$$

where C_2 is the probability limit of $\hat{C}_2$. With $\hat{C}_2$ containing estimated density derivatives up to order 2 and with C_2 containing the corresponding expected derivatives, it can be shown that

$$\sup_{t \,\epsilon\, \mathcal{P}^*} h^2 \left| \left(\hat{C}_2(t) - C_2(t)\right) \right| = h^2 h_p^2 = o\left(N^{-1/2}\right).$$

The results now follows for $\Delta_e(t)$. The argument for $\Delta_1(t)$ is similar.

To analyze the second gradient component, Lemma 5 below shows that the estimated trimming and weight functions in this component may be taken as known.

Lemma 5. With w_i and $\hat{w}_i$ defined as in Lemma G1, for the second gradient component defined above:

$$\begin{aligned} N^{1/2}\hat{G}_2 &\equiv \sqrt{N}\sum \hat{\tau}_i \left[\hat{M}_{oi} - M_{oi}\right] \hat{w}_i/N = N^{1/2}\hat{G}_2^* + o_p(1), \\ \hat{G}_2^* &= \sum \tau_{io} \left[\hat{M}_{oi} - M_{oi}\right] w_i/N. \end{aligned}$$

Proof of Lemma 5. From the definitions of $\hat{G}_2$ and $\hat{G}_2^*$:

$$N^{1/2}\left[\hat{G}_2 - \hat{G}_2^*\right] = N^{1/2}\sum \left[\hat{M}_{oi} - M_{oi}\right] \left[\hat{\tau}_i\hat{w}_i - \tau_{io}w_i\right]/N.$$

The result follows from repeated application of Lemma 1 and a convergence rate for indicators in Klein (1993).

Employing Lemma 5 and earlier results, it is now possible to show that estimated local smoothing parameters may be taken as fixed at values for which the required degree of bias reduction holds. Let $\hat{E}_{1i} \equiv \hat{S}_{ui}^2$ and $\hat{E}_{2i} \equiv \hat{S}_{vi}^2$. Then, using notation introduced earlier, write:

$$\hat{E}_{ki} \equiv \hat{f}_{ki}/\hat{g}_{ki}.$$

Recall that $\hat{f}_{ki}$ and $\hat{g}_{ki}$ each depend on a local smoothing functions evaluated at pilot estimators. Further denote $\bar{f}_{ki}$ and $\bar{g}_{ki}$ as the corresponding quantities with local smoothing functions evaluated at the expectation of the pilot estimators upon which they depend. Then, define:

$$\bar{S}_{ui}^2 \equiv \bar{E}_{1i} \equiv \bar{f}_{1i}/\bar{g}_{1i}; \quad \bar{S}_{vi}^2 \equiv \bar{E}_{2i} \equiv \bar{f}_{2i}/\bar{g}_{2i}.$$

Employing this notation, Lemma 6 below shows that the gradient based on $\hat{S}_{ui}^2$ and $\hat{S}_{vi}^2$ is appropriately close to that based on $\bar{S}_{ui}^2$ and $\bar{S}_{vi}^2$.

Lemma 6 With $\hat{M}_{oi}$ defined as above, obtain $\bar{M}_{oi}$ by replacing $\hat{S}_{ui}^2$ and $\hat{S}_{vi}^2$ with $\bar{S}_{ui}^2$ and $\bar{S}_{vi}^2$ respectively. In $\hat{G}_2$, replace $\hat{M}_{oi}$ with $\bar{M}_{oi}$ to obtain $\bar{G}_2$. Then:

$$\sqrt{N}\left[\hat{G}_2 - \bar{G}_2\right] = o_p(1).$$

Proof of Lemma 6. From Lemma 5 and the definition of gradient components (1A-2A):

$$\sqrt{N}\left[\hat{G}_2 - \bar{G}_2\right] = \rho_o\sqrt{N}\sum_{i=1}^{N}\tau_i\left[\left(\hat{S}_{ui}/\hat{S}_{vi}\right) - \left(\bar{S}_{ui}/\bar{S}_{vi}\right)\right]v_i w_i/N \; + o_p(1).$$

The above difference will depend on the estimation error in conditional variance components for u and v errors. Denoting these terms by $\rho_o\Delta_u$ and $\rho_o\Delta_v$ respectively, each is $o_p(1)$. For the former (the analysis for the latter is identical):

$$\Delta_u \equiv N^{-1/2}\sum_{i=1}^{N} v_i w_i\left(\hat{S}_{ui} - \bar{S}_{ui}\right)/\hat{S}_{vi}.$$

From a Taylor series expansion, Lemma 1, and with $\psi_i \equiv v_i w_i/(2S_{ui}S_{vi})$:

$$\begin{aligned}\Delta_u &= N^{-1/2}\sum_{i=1}^{N}\left[\hat{S}_{ui}^2 - \bar{S}_{ui}^2\right]\psi_i + o_p(1) \\ &= N^{-1/2}\sum_{i=1}^{N}\left[\left(\hat{f}_{1i}/\hat{g}_{1i}\right) - \left(\bar{f}_{1i}/\bar{g}_{1i}\right)\right]\psi_i + o_p(1).\end{aligned}$$

This term involves differentials in both numerator and denominator. As the arguments for both terms are very similar, here we show that the former term is $o_p(1)$. This term is given as:

$$\Delta_f = N^{-1/2}\sum_{i=1}^{N}\left[\hat{f}_{1i} - \bar{f}_{1i}\right](\psi_i/g_{1i}) + o_p(1).$$

In more compact notation, with $Y_j \equiv u_j^2$ and $z_j \equiv I_{ui}(b_o)$, let:

$$\begin{aligned}\hat{L}_j &\equiv \left[\hat{f}^*(z_j)\right]^{1/2};\; \bar{L}_j \equiv \left[\bar{f}^*(z_j)\right]^{1/2} \\ K_{ij} &\equiv Y_j\frac{\hat{L}_j}{h}K\left[\frac{z_i - z_j}{h}\hat{L}_j\right];\; \bar{K}_{ij} \equiv Y_j\frac{\bar{L}_j}{h}K\left[\frac{z_i - z_j}{h}\bar{L}_j\right].\end{aligned}$$

Then:

$$\Delta_f = N^{-1/2}\sum_{i=1}^{N}\left[K_{ij} - \bar{K}_{ij}\right]Y_j(\psi_i/g_{1i}) + o_p(1).$$

Define:

$$\begin{aligned} k\left(\frac{t-z}{h}l\right) &\equiv \left[K\left(\frac{t-z}{h}l\right)+K\left(\frac{t-z}{h}l\right)'\left(\frac{t-z}{h}\right)l\right] \\ b_j &\equiv \frac{1}{2}\bar{f}_j^{*-1/2}Y_j; \quad a_i \equiv \psi_i/g_{1i} \\ \varepsilon_i &\equiv \sum_{j\neq i}\frac{1}{(N-1)\,h}k\left(\frac{z_i-z_j}{h}\bar{L}_j\right)a_i b_j\left(\hat{f}_j-\bar{f}_j\right). \end{aligned}$$

From a Taylor series expansion in $\hat{f}_j^*$ about $\bar{f}_j^*$ and from Lemma 1:

$$\Delta_f = N^{1/2}\sum_{i=1}^{N}\varepsilon_i/N + r,\ r = o_p(1).$$

For the leading term :

$$E\left(\Delta_f - r\right)^2 = \underbrace{\tfrac{1}{N}\textstyle\sum_{r\neq s}\sum E\left(\varepsilon_r\varepsilon_s\right)}_{\mathbf{C}} + \underbrace{\tfrac{1}{N}\textstyle\sum_{i=1}^{N}E\left(\varepsilon_i^2\right)}_{\mathbf{S}}.$$

For the cross-product (**C**) terms:

$$\begin{aligned} \mathbf{C} &= O(N)E\left[\varepsilon_r\varepsilon_s\right] = C_1 + C_2, \\ \mathbf{C_1} &= O(N)E\sum_{j\neq r,s}\frac{1}{(N-1)^2h^2}k\left(\frac{z_r-z_j}{h}\bar{L}_j\right)k\left(\frac{z_s-z_j}{h}\bar{L}_j\right)b_j^2a_ra_s\left(\hat{f}_j-\bar{f}_j\right)^2 \\ \mathbf{C_2} &= O(N)E\sum_k\sum_{l\neq k}\left[\begin{array}{l}\left(\frac{1}{(N-1)h}k\left(\frac{z_r-z_k}{h}\bar{L}_k\right)b_ka_r\left(\hat{f}_k-\bar{f}_k\right)\right) \\ *\left(\frac{1}{(N-1)h}k\left(\frac{z_s-z_l}{h}\bar{L}_l\right)b_la_s\left(\hat{f}_l-\bar{f}_l\right)\right)\end{array}\right]. \end{aligned}$$

Write:

$$\hat{f}_j-\bar{f}_j = \left[\hat{f}_j[r,s]-\frac{N-3}{N-1}\bar{f}_j\right] + \left[O\left(\left[Nh_p\right]^{-1}\right)c_{rs}+O\left(N^{-1}\right)\bar{f}_j\right],$$

where $\hat{f}_j[r,s]$ is obtained by removing the two terms from $\hat{f}_j$ that depend on observations r and s and where c_{rs} depends on observations r and s and is bounded. Then, it can be shown that $\mathbf{C_1} \to 0$.

Turning to $\mathbf{C_2}$, with $r \neq s \neq k \neq l$:

$$\mathbf{C_2} = O(N)E \left[\begin{array}{c} \left(\frac{a_r}{h} k\left(\frac{z_r - z_k}{h}\bar{L}_k\right) \frac{a_s}{h} k\left(\frac{z_s - z_l}{h}\bar{L}_l\right) b_k b_l\right) \\ * \left(\left[\hat{f}_k[r,s] - \frac{N-3}{N-1}\bar{f}_k\right] + \left[O\left(\left[Nh_p\right]^{-1}\right) c_{rs} + O\left(N^{-1}\right)\bar{f}_j\right]\right) \\ * \left(\left[\hat{f}_l[r,s] - \frac{N-3}{N-1}\bar{f}_l\right] + \left[O\left(\left[Nh_p\right]^{-1}\right) c_{rs} + O\left(N^{-1}\right)\bar{f}_j\right]\right) \end{array} \right].$$

The analysis of this term is based on the differential $\left(\hat{f} - \bar{f}\right)$ and on a property of the function k. For this function, note that:

$$\Delta \equiv \int\int \frac{y}{h} k\left(\frac{t-z}{h} l\right) f(y,z)\, dydz = O\left(h^2\right).$$

To establish this rate for Δ, with $\varepsilon \equiv [t-z]/h$:

$$\Delta = -\int y \left[\int k\left[\varepsilon l\left(t - h\varepsilon\right)\right] f\left(t - h\varepsilon \mid y\right) d\varepsilon\right] g_y(y) dy.$$

Take a Taylor series expansion in h about $h = 0$. Since k is an even function, odd moments will vanish. The result then follows because the inner integral of the following expression is zero:

$$-\int y f\left(t \mid y\right) \left[\int k\left[\varepsilon l\left(t\right) t\right] d\varepsilon\right] g_y(y) dy = 0.$$

Employing the above result for the function k and an iterated expectations argument, it can be shown that $\mathrm{C}_2 = O(N)O(h^4)O(h_p^4)$ vanishes under the window conditions. A similar and somewhat simpler argument shows that the $\mathbf{S}$ terms also vanish in probability, which completes the proof.

The gradient component $\bar{\mathrm{G}}_2$ is composed of functions of estimated conditional variances. To analyze this component, Lemma 9 below shows that such functions can be written as U-statistics and analyzed by standard projection arguments. To state and prove this result, we will require some notation, some of which has already been introduced. Recall that $\hat{S}_{ui}^2 \equiv \hat{f}_{1i}/\hat{g}_{1i}$ and that with $k_1[i,j]$ and $k_1^*[i,j]$ as the kernel functions underlying $\hat{m}_{1i}$ and $\hat{g}_{1i}$ respectively:

$$\hat{S}_{ui}^2 \equiv \left[\frac{1}{N-1}\sum_{j\neq i}^{N} u_j^2 k_1[i,j]\right] / \left[\frac{1}{N-1}\sum_{j\neq i}^{N} k_1^*[i,j]\right]. \tag{3A}$$

A similar expression defines $\hat{S}_{vi}^2$. For notational convenience, let $I_{1i} \equiv I_{ui}$ and $I_{2i} \equiv I_{vi}$. Then, with $S_{1i}^2 \equiv E\left(u_i^2 | I_{1i}\right)$ and $S_{2i}^2 \equiv E\left(v_i^2 | I_{2i}\right)$, write:

$$S_{ki}^2 \equiv \left[E_{ki} g_k\right] / g_{ki} \equiv f_{k_i i} / g_{ki}, \ k = 1, 2. \tag{4A}$$

It will also be convenient to write the vector valued weight function, shown in Lemma G1, compactly as:

$$w_i \equiv r\left(X_i\right) + s\left(X_i\right) v_i, \tag{5A}$$

where r and s are vector-valued functions of X_i.

Lemma 7 (U-Statistic Projection). Denote $\nabla_k^d f\left(x_1, x_2\right)$ as the d^{th} derivative of the function f with respect to its k^{th} argument. For (x_1, x_2) bounded and $c > 0$ finite, assume that the function f has bounded derivatives:

$$\left|\nabla_k^d f\left(x_1, x_2\right)\right| \leq c \text{ for } k = 1, 2 \text{ and } d = 1, 2$$

define:

$$\hat{f}_i \equiv f\left(\hat{S}_{ui}^2, \hat{S}_{vi}^2\right); \quad f_i \equiv f\left(S_{ui}^2, S_{vi}^2\right); \ d_{ki} \equiv \nabla_k^1 f\left(S_{ui}^2, S_{vi}^2\right)$$

With τ_{io} as the trimming function evaluated at population quantiles and with the weight function characterized as in 5A, let:

$$\begin{aligned} \alpha_{ki} &\equiv \left[s\left(X_i\right) \tau_{io} d_{ki} / g_{ki}\right] v_i^2 \\ \bar{\alpha}_{ki} &\equiv E\left[\alpha_{ki} | I_{ki}\right], \ k = 1, 2. \end{aligned}$$

Then:

$$\begin{aligned} \Delta &\equiv N^{-1/2} \sum \left[\hat{\tau}_i \hat{f}_i - \tau_i f_i\right] v_i w_i \\ &= \sqrt{N} \bar{\varepsilon}_u^* + \sqrt{N} \bar{\varepsilon}_v^* + o_p\left(1\right), \\ \bar{\varepsilon}_u^* &\equiv \sum \left[u_i^2 - S_{ui}^2\right] \bar{\alpha}_{1i} / N \\ \bar{\varepsilon}_v^* &\equiv \sum \left[v_i^2 - S_{vi}^2\right] \bar{\alpha}_{2i} / N. \end{aligned}$$

Proof of Lemma 7. Employing the same arguments as in Lemma G1, we may take the trimming function as known. Then, from (4A) and with $\hat{S}_{ui}^2 \equiv \hat{f}_{1i} / \hat{g}_1 i$, it follows from Lemma 1 and a Taylor series expansion that:

$$\begin{aligned} \Delta &= \Delta_1 + \Delta_2 + o_p\left(1\right), \\ \Delta_k &\equiv \sqrt{N} \sum \tau_{io} v_i w_i d_{ki} \left[\hat{f}_{ki} / \hat{g}_{ki} - f_{ki} / g_{ki}\right] / N. \end{aligned}$$

Simplifying Δ_1:

$$\begin{aligned}\Delta_1 &= N^{-1/2}\sum \tau_{io}v_iw_id_{1i}\left[\ (\hat{m}_{1i}/\hat{g}_{1i}) - m_{1i}/g_{1i}\right]\left[\ \hat{g}_{1i}/g_{1i}\right] + \\ & \quad N^{-1/2}\sum \tau_{io}v_iw_id_{1i}\left[\ (\hat{m}_{1i}/\hat{g}_{1i}) - m_{1i}/g_{1i}\right]\left[(\hat{g}_i - g_i)\ \ /g_{1i}\right].\end{aligned}$$

From Lemma 1, the second term is $o_p(1)$. Therefore, from (3A) and with $\alpha^*_{1i} \equiv \tau_{io}v_iw_id_{1i}/g_{1i}$:

$$\begin{aligned}\Delta_1 &= N^{-1/2}\sum_i \alpha^*_{1i}\left[\ \hat{m}_{1i} - m_{1i}\hat{g}_i/g_{1i}\right] + o_p(1). \\ &= \frac{N^{1/2}}{N(N-1)}\sum_i\sum_{j\neq i}\rho^*_{ij},\ \rho^*_{ij} \equiv \alpha^*_{1i}\left[u_j^2k_1[i,j] - (m_{1i}/g_{1i})\,k_1^*[i,j]\right] + o_p(1) \\ &= N^{1/2}U_N,\ U_N \equiv \begin{pmatrix} N \\ 2 \end{pmatrix}^{-1}\sum_i\sum_{j>i}\rho_{ij},\ \rho_{ij} = \left(\rho^*_{ij} + \rho^*_{ji}\right)/2.\end{aligned}$$

Since U_N is a U-statistic and $E\left(\alpha^*_{1i}|I_{1i}\right) = \bar{\alpha}_{1i}$, from standard projection arguments:[18]

$$\Delta_1 = \sqrt{N}\bar{\varepsilon}^*_u + o_p(1).$$

Employing the same arguments as above, Δ_2 is similarly characterized as $\sqrt{N}\bar{\varepsilon}^*_v$ and the lemma follows .

Employing the above lemma, it is now possible to characterize the second gradient component.

---

[18] With $Z_i \equiv (Y_{1i}, Y_{2i}, X_i)$, under local smoothing (see Lemmas 2-3), for C(Z_i) bounded and $\varepsilon > 0$:

$$E\left(\rho^*_{ij}|Z_i\right) = N^{-1/2-\varepsilon}C(Z_i) \Rightarrow E\left(\rho^*_{ji}\right) = o\left(N^{-1/2}\right).$$

Therefore:

$$\sqrt{N}U_N - \hat{U}_N = o_p(1),\ \hat{U}_N \equiv \frac{1}{N}\sum_i E\left(\rho^*_{ji}|Z_i\right).$$

The inner expectation has the form:

$$E\left(\rho^*_{ji}|Z_i\right) = \varepsilon^*_{ui} + h^2r_i,\ r_i = O(1),$$

where $E\left(\varepsilon^*_{ui}\right) = 0$. Further:

$$E\left(\rho^*_{ji}\right) = o(N^{-1/2}) \Rightarrow h^2E(r_i) = o\left(N^{-1/2}\right).$$

With $\bar{r} \equiv \Sigma r_i/N$, it can be shown that $\mathrm{Var}\left(\sqrt{N}h^2\bar{r}\right) \to 0$. The result follows.

Lemma G2. The second gradient component has the characterization:

$$\sqrt{N}\bar{G}_2 = \sqrt{N}\bar{\varepsilon}_2,$$

where the ε_{2i} are i.i.d. with expectation 0 and finite variance.

Proof of Lemma G2 . From Lemmas 5-8, we may evaluate $\sqrt{N}\hat{G}_2$ as the true weights and with expected densities replacing estimated densities in the local smoothing functions. Employing the same arguments as in Lemma G1, we will also be able to take the trimming function as known. Therefore:

$$\begin{aligned}\sqrt{N}\hat{G}_2 &= \rho_o\sqrt{N}\sum_{i=1}^{N}\tau_{io}\left[\hat{S}_{ui}/\hat{S}_{vi} - S_{ui}/S_{vi}\right]v_i w_i \\ &\equiv \rho_o\sqrt{N}\sum_{i=1}^{N}\tau_{io}\left[f\left(\hat{S}^2_{ui},\hat{S}_{vi},v_i w_i\right) - f\left(S^2_{ui},S^2_{vi},v_i w_i\right)\right].\end{aligned}$$

The proof now immediately follows from Lemma 9 with $\bar{\varepsilon}_2 \equiv \bar{\varepsilon}^*_u + \bar{\varepsilon}^*_v$.

7.2 Main Results

Recall that the third gradient component for the second stage estimator depends on $\hat{\eta}$, the estimator for the nuisance parameter vector from the Y_1-model. To analyze such first-stage estimation uncertainty, Theorem 1 below characterizes the components of $\hat{\eta}$.

Theorem 1: First Stage Consistency and Characterization. Define:

$$v_i^2(\pi) \equiv (Y_{2i} - X_i\pi)^2,$$

where

$$E\left[v_i^2(\pi_o) \mid I_{vi}(\delta_o)\right] = E\left[Y_i(\pi) \mid X_i\right].$$

Define

$$\begin{aligned}\hat{R}(\delta;\pi) &\equiv \frac{1}{N}\sum_{i=1}^{N}\hat{\tau}_i\hat{r}_i^2(\delta;\pi),\ \hat{r}_i(\delta;\pi) \equiv v_i^2(\pi) - \hat{E}\left[v_i^2(\pi) \mid I_{vi}(\delta)\right] \\ \hat{\delta}(\pi) &= \arg\min_{\delta}\hat{R}(\delta;\pi), \\ \hat{S}_{v_i}(\pi) &\equiv \left[E\left((Y_{2i} - X_i\pi)^2 | I_{vi}\left(\hat{\delta}(\pi)\right)\right)\right]^{1/2}.\end{aligned}$$

With $\hat{\pi}_{ols}$ as the OLS estimator for π_o, let:

$$\hat{X}_i^* \equiv X_i/\hat{S}_{v_i}\left(\hat{\pi}_{ols}\right);\ X_i^* \equiv X_i/S_{v_i}\left(\pi_o\right).$$

Then, with $\Omega \equiv p\lim\left(X^{*'}X^*/N\right)$ and with $\varepsilon_{\pi i} \equiv X_i^{*\prime}v_i$, the GLS estimator of π, $\hat{\pi}$, satisfies:

$$1)\ \sqrt{N}\left[\hat{\pi} - \pi_o\right] = \Omega^{-1}\sqrt{N}\sum \varepsilon_{\pi i}/N + o_p\left(1\right).$$

Define

$$\begin{aligned} R\left(\delta;\pi\right) &\equiv \frac{1}{N}\sum_{i=1}^{N}\tau_{io}r_i^2\left(\delta;\pi\right),\ r_i\left(\delta;\pi\right) \equiv v_i^2\left(\pi\right) - E\left[v_i^2\left(\pi\right)\mid I_{vi}\left(\delta\right)\right] \\ w_i &\equiv 2\tau_{io}\frac{\partial}{\partial\delta}r_i\left(\delta_o;\pi_o\right),\ w_i^* \equiv w_i - E\left[w_i|I_{vi}\left(\delta_o\right)\right] \\ R_{11} &\equiv p\lim\left[\frac{\partial^2}{\partial\delta\partial\delta'}R\left(\delta_o;\pi_o\right)\right];\ R_{21} \equiv p\lim\left[\frac{\partial^2}{\partial\pi\partial\delta'}R\left(\delta_o;\pi_o\right)\right]. \end{aligned}$$

The estimator for the index parameters, $\hat{\delta}\left(\hat{\pi}\right)$, satisfies:[19]

$$2)\ \sqrt{N}\left[\hat{\delta} - \delta_o\right] = -R_{11}^{-1}\sqrt{N}\left[\sum r_i\left(\delta_o;\pi_o\right)w_i^*/N + R_{21}\left[\hat{\pi} - \pi_o\right]\right] + o_p\left(1\right).$$

Proof of Theorem 1. For (1), the proof is immediate. For (2), accounting for estimation uncertainty in $\hat{\pi}$, the proof follows an adaptation of Ichimura (1993) under local smoothing or from an application of the arguments used in Theorem 3 below.[20]

Recall from (1A-2A) at the beginning of the Appendix that the second stage estimator has three gradient components, with the first two being characterized in Lemmas G1-2 above. Lemma G3 characterizes the third gradient component.

[19] This characterization holds under a more general semiparametric formulation of the Y_2-model. Here, to emphasize identification issues, we have focused on the case where the Y_2 model is linear with an unknown conditional variance function.

[20] With the weight redefined for the second-stage estimator, first and second-stage gradients have a similar structure. The intermediate lemmas used to prove Theorem 3 could also then be employed to prove Theorem 1 under estimated local smoothing.

Lemma G3. Referring to (1A-2A), let $Q_1 \equiv p\lim\left(\nabla G_\eta\left(\theta_o, b_o; \eta_o\right)\right)$. Employing the notation Theorem 1 above, define:

$$\varepsilon_{3i} \equiv Q_1 \left[\begin{array}{c} \Omega^{-1}\varepsilon_{\pi i} \\ \left(-R_{11}^{-1}\left[r_i\left(\delta_o;\pi_o\right)w_i^* + R_{21}\Omega^{-1}\varepsilon_{\pi i}\right]\ \right) \end{array} \right].$$

Then:

$$\sqrt{N}\hat{G}_3 = \sqrt{N}\bar{\varepsilon}_3 + o_p(1),\ \bar{\varepsilon}_3 \equiv \sum_{i=1}^{N} \varepsilon_{3i}/N.$$

Proof of Lemma G3. The proof follows from (1A-2A), Theorem 1, and a standard uniform convergence result.

Theorem 2: Second Stage Consistency and Identification. Let Z be the matrix with i^{th} row:

$$\left[W_i \quad \left(\ S_{ui}\left(\theta_o, b_o\right)/S_{vi}\right)v_i\right].$$

Then, the model is identified under the constant correlation assumption if Z has full column rank.

Proof of Theorem 2. Recall from (D5) that: with $\alpha \equiv (\theta, \rho, b)$:

$$\begin{array}{rcl} W & \equiv & \left[X \quad Y_2\right],\ S_{ui}\left(\theta, b\right) \equiv E\left[\left(Y_{1i} - W_i\theta\right)^2 | I_{ui}\left(b\right)\right]^{1/2} \\ M\left(\alpha\right) & \equiv & W\theta + \rho S_u\left(\theta, b\right) \equiv M;\ M\left(\theta_o, \rho_o, b_o\right) \equiv M_o \\ \hat{\alpha} & \equiv & \arg\hat{S}\left(\alpha\right) = \arg\min\hat{Q}\left(\alpha\right),\ \hat{Q}\left(\alpha\right) \equiv \hat{S}\left(\alpha\right) - \hat{S}\left(\alpha_o\right). \end{array}$$

Replace all estimated functions $\hat{Q}\left(\alpha\right)$ with their uniform probability limits to obtain $Q\left(\alpha\right)$, It can be shown that $\hat{Q}\left(\alpha\right) - Q\left(\alpha\right)$ is, uniformly in α, $o_p(1)$. Further, the function $Q\left(\alpha\right)$ converges uniformly in the parameters to its expectation given as: $E[\Delta_{1i} + \Delta_{2i}]^2$,

$$\begin{array}{rcl} \Delta_{1i} & \equiv & W_i\left(\theta - \theta_o\right) \\ \Delta_{2i} & \equiv & \left[\rho S_{ui}\left(\theta, b\right) - \rho_o S_{ui}\left(\theta_o, b_o\right)\right]v_i/S_{vi}. \end{array}$$

With $\left(\Delta_{1i} + \Delta_{2i}\right)^2$ minimized at the true parameter values, consistency will follow if this minimum is unique. If the minimum is not unique, it must be

the case that $\Delta_{1i} + \Delta_{2i} = 0$ at all potential minimizing parameter values. Then:

$$W_i(\theta - \theta_o) + [\rho S_{ui}(\theta, b) - \rho_o S_{ui}(\theta_o, b_o)] v_i / S_{vi} = 0, \tag{A}$$

from which it follows that:

$$\begin{aligned} \rho^2 S_{ui}^2(\theta, b)\left(v_i^2/S_{vi}^2\right) &= \rho_o^2 S_{ui}^2(\theta_o, b_o)\left(v_i^2/S_{vi}^2\right) \\ &\quad -2\rho_o S_{ui}(\theta_o, b_o)(v_i/S_{vi}) W_i(\theta - \theta_o) \\ &\quad +(\theta - \theta_o)' W_i' W_i(\theta - \theta_o). \end{aligned}$$

Taking an expectation conditioned on X_i :

$$\begin{aligned} \rho^2 S_{ui}^2(\theta, b) &= \rho_o^2 S_{ui}^2(\theta_o, b_o) - 2\rho_o S_{ui}(\theta_o, b_o) S_{vi}(\theta_2 - \theta_{2o}) \\ &\quad +(\theta - \theta_o)' E[\, W_i' W_i \mid X_i](\theta - \theta_o). \end{aligned} \tag{B}$$

From the definition of $S_{ui}^2(\theta, b)$:

$$\begin{aligned} S_{ui}^2(\theta, b) &= E\left[(Y_i - W_i\theta)^2 \mid X_i\right] = E\left[(\, u_i - W_i(\theta - \theta_o))^2 \mid X_i\right] \\ &= S_{ui}^2(\theta_o, b_o) - 2E(u_i v_i | X_i)(\theta_2 - \theta_{2o}) + (\theta - \theta_o)' W_i' W_i(\theta - \theta_o) \\ &= S_{ui}^2(\theta_o, b_o) - 2\rho_o S_{ui}(\theta_o, b_o) S_{vi}(\theta_2 - \theta_{2o}) + (\theta - \theta_o)' W_i' W_i(\theta - \theta_o). \end{aligned} \tag{C}$$

With $\alpha \equiv \left[\left(1 - \rho_o^2\right)/\left(1 - \rho^2\right)\right]^{1/2}$, differencing the expressions in (B) and (C):

$$\begin{aligned} \rho^2 S_{ui}^2(\theta, b) - S_{ui}^2(\theta, b) &= \rho_o^2 S_{ui}^2(\theta_o, b_o) - S_{ui}^2(\theta_o, b_o) \\ &\Leftrightarrow \quad S_{ui}(\theta, b) = \alpha S_{ui}(\theta_o, b_o). \end{aligned} \tag{D}$$

Substituting (D) into (A):

$$\begin{aligned} W_i(\theta - \theta_o) + (\rho\alpha - \rho_o) S_{ui}(\theta_o, b_o) v_i / S_{vi} &= 0 \Leftrightarrow \\ [W_i\ ,\ (\ S_{ui}(\theta_o, b_o)/S_{vi})\, v_i] \begin{bmatrix} \theta - \theta_o \\ \rho\alpha - \rho_o \end{bmatrix} &= 0. \end{aligned}$$

Under a full rank assumption, $\theta = \theta_o$ and $\rho\alpha = \rho_o$. Since $\theta = \theta_o$, from (C), $S_{ui}(\theta_o, b) = S_{ui}(\theta_o, b_o)$. Consequently, from (D), $\alpha = 1$. With $\rho\alpha = \rho_o$, it follows that $\rho = \rho_o$. As shown above, $S_{ui}^2(\theta_o, b) = S_{ui}^2(\theta_o, b_o)$. It can now also readily be shown that $b = b_o$ when S_{ui} is subject to a single index assumption (Ichimura 1993).

Theorem 3 : Asymptotic Normality of the Second Stage Estimator. Define ε_{ki}, $k = 1, 2, 3$, as in Lemmas G1-3 and let:

$$\varepsilon_i \equiv \varepsilon_{1i} + \varepsilon_{2i} + \varepsilon_{3i}.$$

From (1A) of the previous section and with $H_o \equiv E\left[H\left(\alpha_o; \eta_o\right)\right]$:

$$\sqrt{N}\left[\hat{\alpha} - \alpha_o\right] \xrightarrow{d} Z,\ Z \tilde{}\ N\left(0, H_o^{-1} E\left(\varepsilon_i \varepsilon_i'\right) H_o^{-1}\right).$$

Proof of Theorem 3. With $\alpha^+ \epsilon\ (\hat{\alpha}, \alpha_o)$, from (1A-3A) and standard Taylor Series Arguments:

$$\sqrt{N}\left[\hat{\alpha} - \alpha_o\right] = -\left[\hat{H}\left(\alpha^+; \hat{\eta}\right)\right]^{-1}\left[\sqrt{N}\left(\hat{G}_1 + \hat{G}_2 + \hat{G}_3\right)\right],$$

For the Hessian term, from standard uniform convergence arguments: $\hat{H} \xrightarrow{p} H_o$. For the gradient, from Lemmas G1-3:

$$\sqrt{N}\left(\hat{G}_1 + \hat{G}_2 + \hat{G}_3\right) = \sqrt{N}\bar{\varepsilon},\ \bar{\varepsilon} = \sum_{i=1}^{N} \varepsilon_i / N.$$

The result now follows.

References

[1] Abramson, I.S. (1982): "Bandwidth Variation in Kernel Estimates- A Square Root Law," *The Annals of Statistics,* 10, 1217-1223.

[2] Bollerslev, T. (1990): " Modelling the Coherence in Short-Run Nomial Exchange Rates: A Multivariate Generalized GARCH Approach," *Review of Economics and Statistics, 72,* 498-505.

[3] Dagenais, M., and D.Dagenais (1997): "Higher Moment Estimators for Linear Regression Models with Errors in Variables," *Journal of Econometrics, 76 (1-2),* 193-222.

[4] Hall, P. (1990): "On the Bias of Variable Bandwidth Curve Estimators," *Biometrika*, 77, 529-535.

[5] Ichimura, H, (1993):"Semiparametric least squares (SLS) and weighted SLS estimation of single index models" *Journal of Econometrics*, 58, 71-120.

[6] Klein, R. (1993): "Specification Tests for Binary Choice Models Based on Index Quantiles, *Journal of Econometrics,* 59, 343-375.

[7] Klein, R. and F.Vella (2004): "A Semiparametric Model for Binary Response and Continuous Outcomes Under Index Heteroscedasticity," unpublished manuscript.

[8] Lewbel, A. (1997): "Constructing Instruments for Regressions with Measurement Error when No Additional Data are Available, With an Application to Patents and R&D," *Econometrica*, 65, 1201-1213.

[9] Lewbel, A. (2004): "Identification of Heteroskedastic Endogenous Models or Mismeasured Regressor Models," unpublished manuscript.

[10] Newey, W. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, v. 4, Chapter 36, Amsterdam, North Holland.

[11] Pakes, A. and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027-1058.

[12] Powell, J.L.,J.H. Stock, and T.M. Stoker (1989):"Semiparametric Estimation of Weighted Average Derivatives," *Econometrica*, 57, 1403-1430.

[13] Rigobon, R, (2003): "Identification through heteroscedasticity," *Review of Economics and Statistics,* 85, 777-792.

[14] Rummery, S., F.Vella and M.Verbeek (1999):"Estimating the Returns to Education for Australian Youth via Rank-Order Instrumental Variables," *Labour Economics,* 6, 491-507.

[15] Serfling, R.S. (1980) : *Approximation Theorems of Mathematical Statistics.* New York; Wiley.

[16] Sentana, E. and G.Fiorentini (2001), "Identification, Estimation and Testing of Conditional Heteroskedastic Factor Models,"*Journal of Econometrics,* 102, 143-164.

[17] Silverman, P. (1986): *Density Estimation.* New York; Chapman and Hall.

[18] Staiger, R. and J.Stock (1999): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 68, 1055-1096.

[19] Vella, F. and M.Verbeek (1997): "Rank Order as an Instrumental Variable" unpublished manuscript.